Lecture reviews — Week 07

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

Week 7 keypoints

- supervised/unsupervised
- preprocessing is key
- baseline methods:
  - classification: Naive Bayes, (Logistic regression,) KNN
  - clustering: K-means, dendrograms
  - dim. reduction: PCA, UMAP
- don’t forget evaluation keypoints (see lesson 2)

Weeks 2–3 keypoints

- Evaluation methodology (see the corresponding slide)
- How to assess the quality of a corpus
- How to assess the quality of a system
- How to assess the quality of an evaluation

NLP evaluation protocol

from a design by Layerace / Freepik

1. Define a control task
2. Produce a reference (golden truth)
3. Assess the quality of the reference
4. Evaluate NLP system(s) on the reference
5. Compare evaluations (statistical significance)
6. Publish and discuss results

Week 7 – study case

Some financial company offers you to work on
“*fraud detection using Natural Language Technology applied to client documents*”.

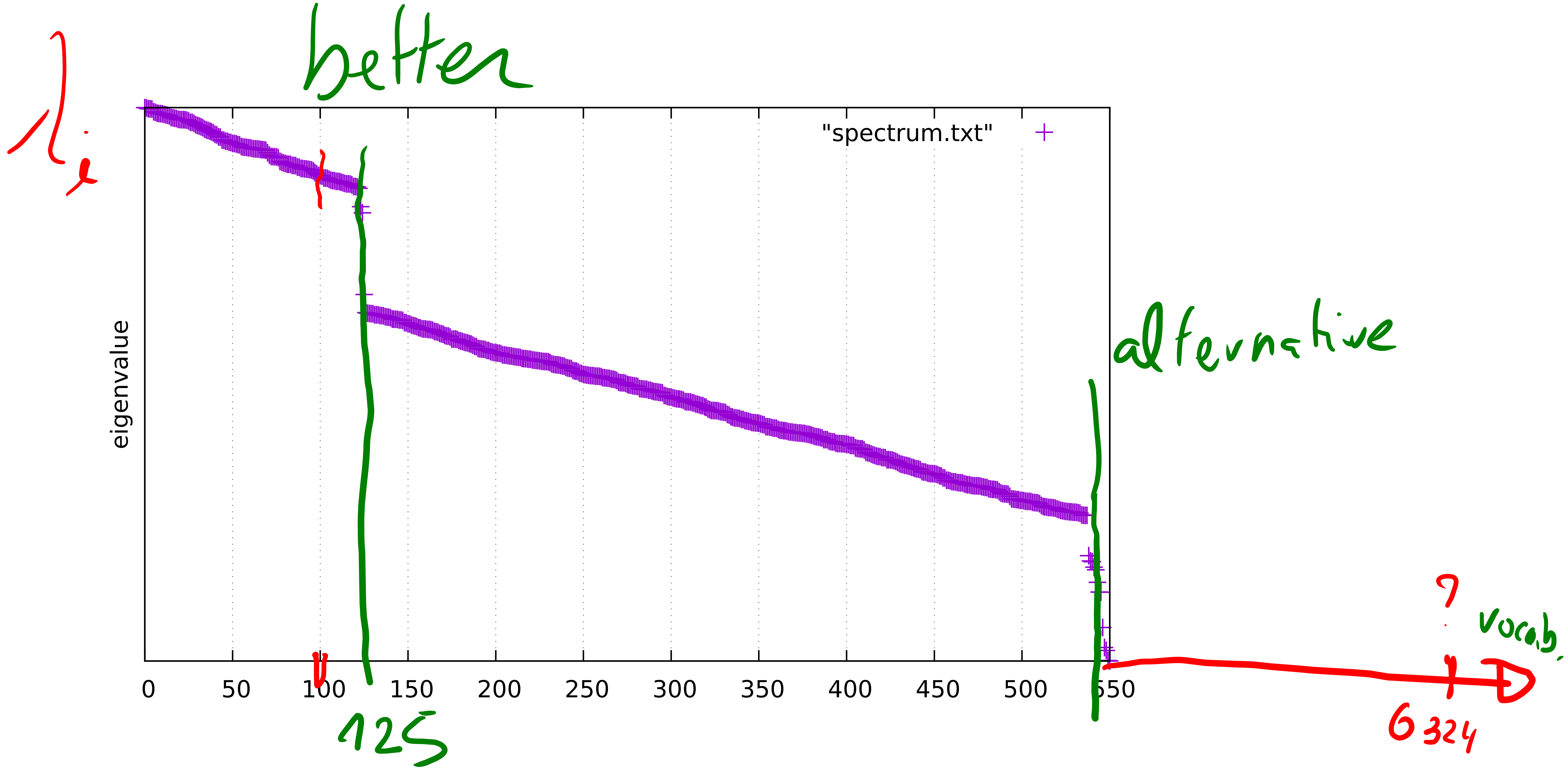
① Some preliminary work has already been performed by a former intern who created document vectors based on an indexing set of 6’324 terms and reduced them to vectors of size 100 using PCA.

Reviewing his/her work and report, you found a graph related to the corresponding singular values.
Next slide shows a (rescaled) zoom on the first 550 left-most points in that graph.

Week 7 – study case

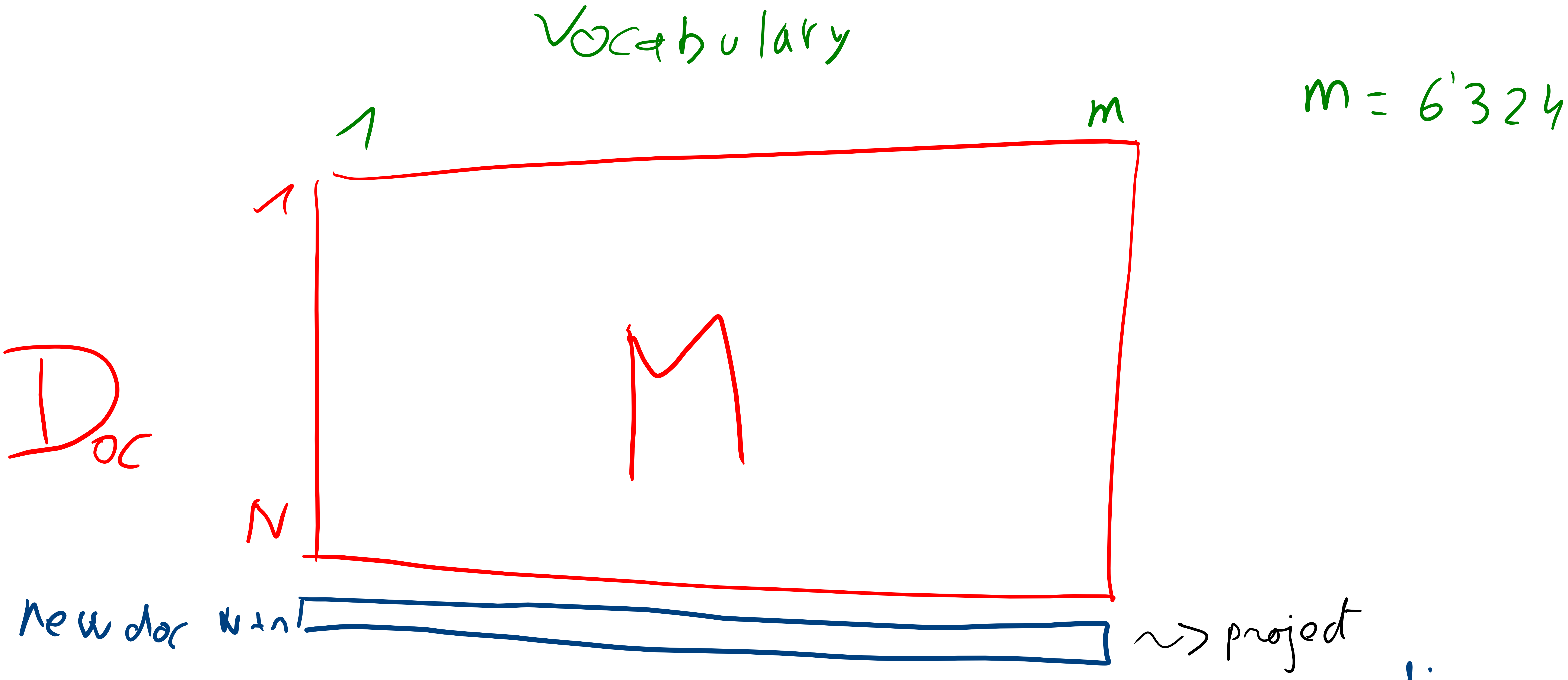
$$\frac{\lambda_i}{\sum \lambda_i}$$

$$\sum_{j=1}^{i} \lambda_j$$



a) What is the abscissa (x-value, horizontal axis) of the right-most point in the original complete graph (not reported here)?

b) What do you think about the intern's methodology for selecting the dimension of the vector space? Would you have performed differently? If yes, how?



N documents

$$M^t M \simeq U_q^t \overset{\substack{\text{diag}\\ \uparrow}}{\Lambda_q} V_q$$

Week 7 – study case

$$\underset{class}{\text{Argmax}}\ P(class|doc)$$

vectorize documents

↳ $P(doc|class) = \prod P(word|class)$

② Before considering more sophisticated Deep-Learning methods, you wisely decide to start with a simple baseline, namely a Naive Bayes model (on the former representation).

a) Based on your former answer, what is the input of the Naive Bayes module?
What is the output?
What are the parameters?
What is needed for training such a model?

set of vectors

$P(Class) \cdot P(doc|class)$

b) Concretely, what probability should be computed as an output from the (very simple excerpt of) client document:

My salary is about 10'000 CHF and I don't pay any tax.

$$\prod_{i=1}^{125} P(\text{word vector}|class)$$

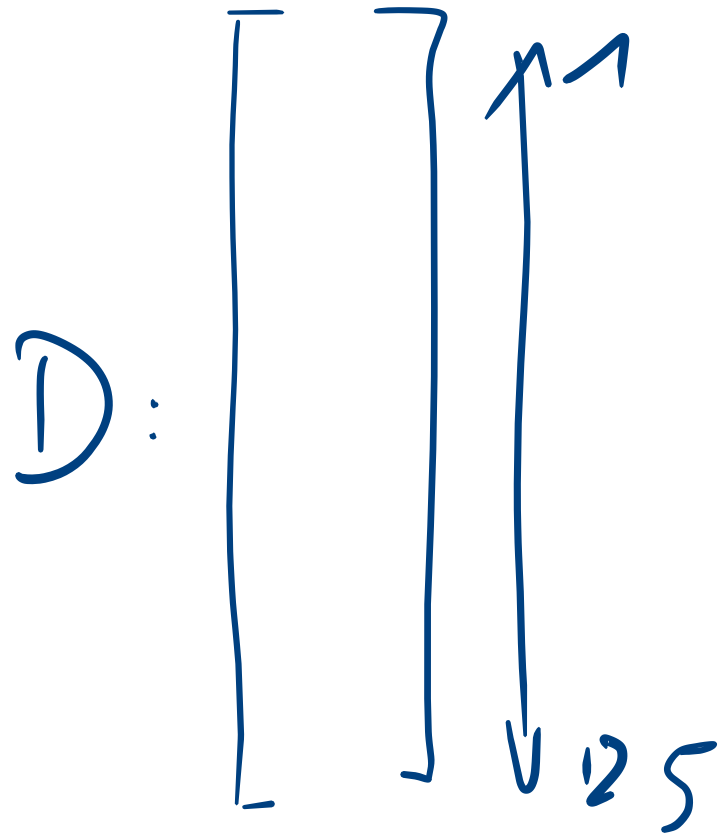
parameters

$P(class)$ for $\begin{cases} class = true \\ class = false \end{cases}$

$P(\text{word vector}|class)$ for all

Doc : My salary . . . .

↳ project D on 125 eigenvectors



for class=true:

$$P(\text{true}) \cdot \prod_{i=1}^{125} P(f_i \mid \text{true})$$

Week 7 – study case

③ From your first analysis of the baseline results, you realize that single tokens do not adequately capture dependencies that clearly appear at the syntactic level (for instance the one between “*don’t*” and “*pay*” in the former example).
Using some syntactic parser, you are able to transform the former example sentence

My salary is about 10’000 CHF and I don’t pay any tax.

into:

SALARY-10K-RANGE not_pay tax

a) What probability would then be computed as the resulting output by the Naive Bayes model in such a case?

b) Compared to former Naive Bayes model, what is the main fundamental reason why you can reasonably expect the results to be better?

$$P(class) \cdot P(SALARY\ldots \mid class) \cdot P(not.pay \mid class) \cdot P(tax \mid class)$$

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE – LAUSANNE
POLITECNICO FEDERALE – LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY – LAUSANNE

Faculté Informatique et Communication
Introduction to Natural Language Processing (Ms; CS-431)
Chappelier, J.-C. & Rajman, M.

CS-431 Hands On Text Classification

J.-C. Chappelier M. Rajman

v. 2021118 – 1

QUESTION I [3 pt]

(from Fall 2018 quiz 4)

The Naïve Bayes algorithm is used in the framework of a sentiment analysis application to determine, for any input tweet, which, among a predefined set of sentiments, best corresponds to the mood expressed in the tweet.

Does the performed tweet classification task have to be supervised in this case?

[✓] yes [ ] no [ ] it depends on the implementation

Let us assume that only two sentiments are considered ("joyful" and "sad") and that typically 70% of the tweets are "joyful".

$P(\text{joy})$ $\Longrightarrow P(\text{sad}) = 30\%$

$P(\text{class}) \cdot \underbrace{P(\text{doc} \mid \text{class})}_{\prod P(w \mid \text{class})}$

To which sentiment would the Naïve Bayes algorithm associate a tweet indexed by only two terms w_1 and w_2, if:

$\to P(w_1 \mid \text{joy})$ $P(w_1 \mid \text{sad})$

- 10% of the occurrences of indexing terms in "joyful" tweets and 20% of the occurrences of indexing terms in "sad" tweets are w_1 ; while
- 30% of the occurrences of indexing terms in "joyful" tweets and 25% of the occurrences of indexing terms in "sad" tweets are w_2?

$\to 70\% \cdot 10\% \cdot 30\%$

[ ] sad [✓] joyful [ ] undecidable

$\hookrightarrow 30\% \cdot 20\% \cdot 25\%$

QUESTION II [2 pt]

(from Fall 2017 quiz 4)

Consider the following matrix of measures over a set of three items:

$$\begin{array}{c|ccc} & a & b & c \\ \hline a & 0 & 5 & 2 \\ b & 5 & 0 & 2 \\ c & 2 & 2 & 0 \end{array}$$

$d(a\ b) > d(a,c) + d(c,b)$

What type(s) of measure is this matrix compatible with?

[✓] A dissimilarity only.

[ ] A dissimilarity and a distance/metric.

NO!

[ ] None of the two

QUESTION III [4 pt]

(from Exam 2019)

You're working on an email classification software (and have some corpus).

In order to better understand your corpus, you plan to cluster it using dendrograms. To do so:

- you represent each email body by the empirical probability distribution over the tokens it contains (simply estimated by their relative frequencies);
- and make use of the Hellinger distance.

→ euclidian($\sqrt{\text{vector}}$)

D

dif: $em_1 - em_2$

What is the distance between the following two email bodies:

email 1: *ski sun money sun*

email 2: *sun ibm sun apple money sun money sun*

	ski	sun	money	ibm	apple	
email 1	1/4	2/4	1/4	0	0	→ sum = 4 → proba dist
email 2	0	4/8	2/8	1/8	1/8	→ sum = 8
diff :	1/4	0	0	1/8	1/8	→ $\frac{1}{\sqrt{2}}$

QUESTION IV [3 pt]

(from Exam 2019)

You run the dendrogram clustering algorithm using complete linkage. At some point, it reaches a state where what remains to be clustered are the two clusters, G_1 and G_2, that have already been build so far, and two email bodies, B_1 and B_2. Here are the distances between each of them:

	~~B_1~~	B_2	G_1	~~G_2~~
~~B_1~~	0	0.7	0.6	0.2
B_2	0.7	0	0.5	0.3
G_1	0.6	0.5	0	0.4
~~G_2~~	0.2	0.3	0.4	0

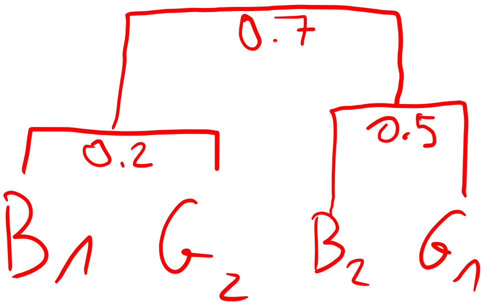
1) Choose min

2) recompute new dist

dist among sets:

- min
- max
- average

Draw the dendrogram corresponding to the final clustering.



	$(B_1 G_2)$
B_2	0.7
G_1	0.6